1. Importing Data and packages

*# Importing pandas, numpy and matpltlib package*

**import pandas as pd**

**import numpy as np**

**import matplotlib.pyplot as plt**

!

!

*# Importing CSV files*

!

!

**dset=pd.read\_csv("general\_data.csv")**

!

!

*# checking first 5 lines of imported file*

!

!

**dset.head()**

Out[5]:

Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager

0 51 No ... 0 0

1 31 Yes ... 1 4

2 32 No ... 0 3

3 38 No ... 7 5

4 32 No ... 0 4

[5 rows x 24 columns]

!

!

*# Columns of imported file*

!

**dset.columns**

Out[6]:

Index(['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',

'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',

'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',

'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',

'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',

'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager'],

dtype='object')

!

*# Data of of imported file*

!

**dset**

Out[8]:

Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager

0 51 No ... 0 0

1 31 Yes ... 1 4

2 32 No ... 0 3

3 38 No ... 7 5

4 32 No ... 0 4

... ... ... ... ...

4405 42 No ... 0 2

4406 29 No ... 0 2

4407 25 No ... 1 2

4408 42 No ... 7 8

4409 40 No ... 3 9

[4410 rows x 24 columns]

1. Treating Data

*# Data treating data by removing null values, duplicate data.*

**dset.isnull()**

Out[9]:

Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager

0 False False ... False False

1 False False ... False False

2 False False ... False False

3 False False ... False False

4 False False ... False False

... ... ... ... ...

4405 False False ... False False

4406 False False ... False False

4407 False False ... False False

4408 False False ... False False

4409 False False ... False False

[4410 rows x 24 columns]

!

!

!

**dset.duplicated()**

Out[11]:

0 False

1 False

2 False

3 False

4 False

4405 False

4406 False

4407 False

4408 False

4409 False

Length: 4410, dtype: bool

!

!

**dset.drop\_duplicates()**

Out[12]:

Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager

0 51 No ... 0 0

1 31 Yes ... 1 4

2 32 No ... 0 3

3 38 No ... 7 5

4 32 No ... 0 4

... ... ... ... ...

4405 42 No ... 0 2

4406 29 No ... 0 2

4407 25 No ... 1 2

4408 42 No ... 7 8

4409 40 No ... 3 9

[4410 rows x 24 columns]

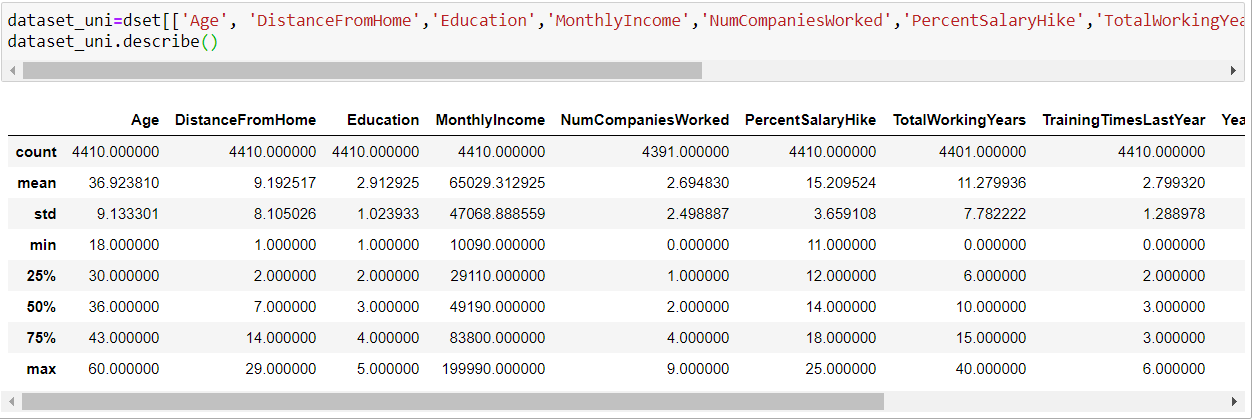
1. Univariate Analysis

**dataset\_uni=dset[['Age','DistanceFromHome','Education','MonthlyIncome','NumCompaniesWorked','PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear','YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']]**

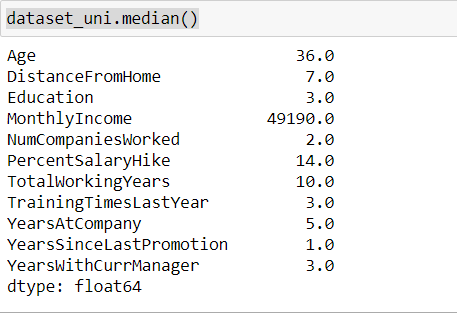
!

!

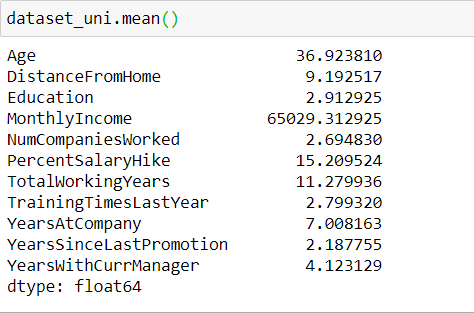
**dataset\_uni.describe()**

****

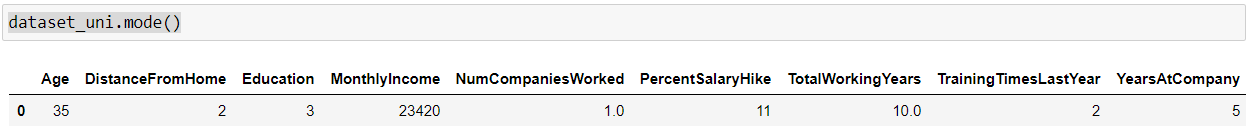
**dataset\_uni.median()**

****

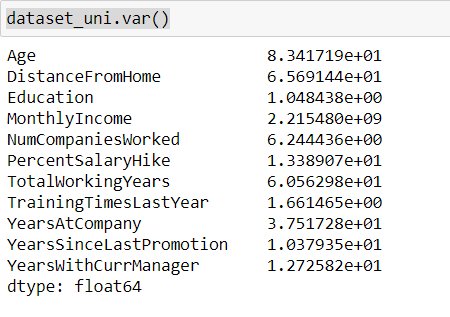
**dataset\_uni.mean()**

****

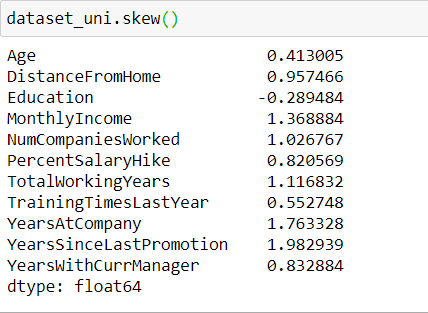
**dataset\_uni.mode()**



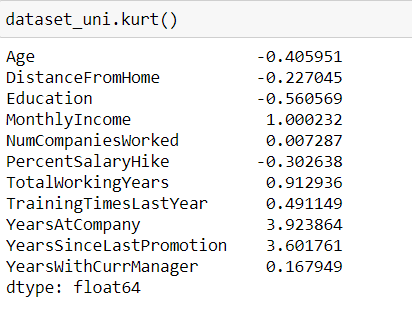
**dataset\_uni.var()**

****

**dataset\_uni.skew()**

****

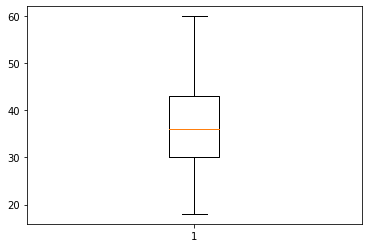
**dataset\_uni.kurt()**

****

* Mean is higher than median so **skewness** is **positive** here
* Here Age, distance from home, education & salary hike is **platykurtic,** all other are **leptokurtic**
* Standard deviation and IQR high for mean monthly income, suggesting attrition across all income bands
* IQR for age is 13years it looks like normal distribution

**box\_plot=dataset\_uni.Age**

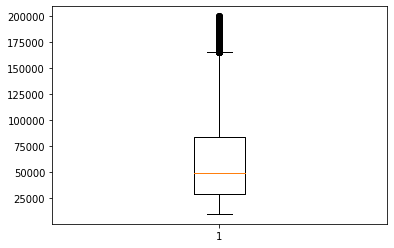
**plt.boxplot(box\_plot)**

****

Age is normally distributed & no outliers

**box\_plot=dataset\_uni.MonthlyIncome**

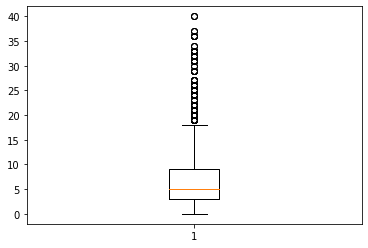
**plt.boxplot(box\_plot)**

****

Monthly income has right skewness & several outliers

**box\_plot=dataset\_uni.YearsAtCompany**

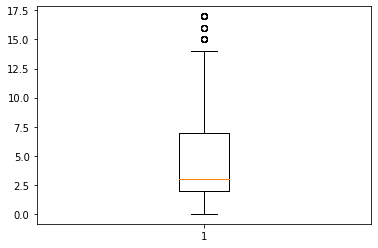
**plt.boxplot(box\_plot)**

****

Years at company have right skewness & several outliers

**box\_plot=dataset\_uni.YearsWithCurrManager**

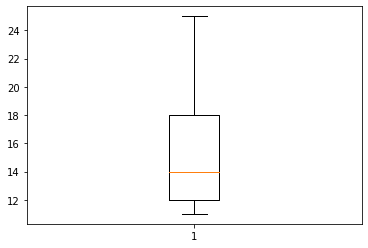
**plt.boxplot(box\_plot)**

****

Years at company have right skewness & few outliers

**box\_plot=dataset\_uni.PercentSalaryHike**

**plt.boxplot(box\_plot)**

****

Salary hike have right skewness & no outliers

1. Non-Parametric Statistical Test (Mann-Whitney)

**import pandas as pd**

**import numpy as np**

**import matplotlib.pyplot as plt**

**dset=pd.read\_csv("C:/Users/psing/Box/LetsUpgrade AI-ML/Day 7 - Numpy Dinesh/Assignment/general\_data.csv")**

**dset.isnull()**

**dset.duplicated()**

**dset.drop\_duplicates()**

**!**

**!**

**dset.isnull().sum() ## to find out null values**

Age 0

Attrition 0

BusinessTravel 0

Department 0

DistanceFromHome 0

Education 0

EducationField 0

EmployeeCount 0

EmployeeID 0

Gender 0

JobLevel 0

JobRole 0

MaritalStatus 0

MonthlyIncome 0

NumCompaniesWorked 19

Over18 0

PercentSalaryHike 0

StandardHours 0

StockOptionLevel 0

TotalWorkingYears 9

TrainingTimesLastYear 0

YearsAtCompany 0

YearsSinceLastPromotion 0

YearsWithCurrManager 0

dtype: int64

**!**

**!**

**dset.mean()**

Age 36.923810

Attrition 0.161224

BusinessTravel 1.607483

Department 1.260544

DistanceFromHome 9.192517

Education 2.912925

EducationField 2.247619

EmployeeCount 1.000000

EmployeeID 2205.500000

Gender 0.600000

JobLevel 2.063946

JobRole 4.458503

MaritalStatus 1.097279

MonthlyIncome 65029.312925

NumCompaniesWorked 2.694810

PercentSalaryHike 15.209524

StandardHours 8.000000

StockOptionLevel 0.793878

TotalWorkingYears 11.279937

TrainingTimesLastYear 2.799320

YearsAtCompany 7.008163

YearsSinceLastPromotion 2.187755

YearsWithCurrManager 4.123129

dtype: float64

**!**

**!**

**# replaced null values with mean**

**dset['NumCompaniesWorked'] = dset['NumCompaniesWorked'].fillna(2.69)**

**dset['TotalWorkingYears'] = dset['TotalWorkingYears'].fillna(11.28)**

**!**

**!**

**#Using sklearn changing/encoding string values to numbers**

**from sklearn import preprocessing**

**le = preprocessing.LabelEncoder()**

**dset['Attrition'] = le.fit\_transform(dset['Attrition'])**

**dset['BusinessTravel'] = le.fit\_transform(dset['BusinessTravel'])**

**dset['EducationField'] = le.fit\_transform(dset['EducationField'])**

**dset['Gender'] = le.fit\_transform(dset['Gender'])**

**dset['Department'] = le.fit\_transform(dset['Department'])**

**dset['MaritalStatus'] = le.fit\_transform(dset['MaritalStatus'])**

**dset['JobRole'] = le.fit\_transform(dset['JobRole'])**

**!**

**!**

**dset1=dset[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',**

**'Education', 'Gender', 'JobLevel', 'JobRole', 'MonthlyIncome',**

**'NumCompaniesWorked', 'PercentSalaryHike', 'TotalWorkingYears', 'TrainingTimesLastYear',**

**'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']]**

**!**

**!**

**# new separate data for attrition & non attrition**

**dset1\_att\_y=dset1[dset1['Attrition']==1]**

**dset1\_att\_n=dset1[dset1['Attrition']==0]**

**!**

**!**

**!**

**!**

**!**

**from scipy.stats import mannwhitneyu**

**num\_cols=list(dset1.columns) -----------------🡪 to list number of columns**

**for i in num\_cols:**

**print(f"Test {i} vs Attrition: \n")**

**print(f"\tH0: attrited and non-attrited has no significant differences in the {i}")**

**print(f"\tH1: attrited and non-attrited has significant differences in the {i}")**

**stats, p = mannwhitneyu(dset1\_att\_y[i],dset1\_att\_n[i])**

**print("\n\t Value of P is ",p, " & decision is as following\n")**

**if p < 0.05:**

**print(f"\t value of less then 0.05, H0 Rejected & H1 Accepted")**

**else:**

**print(f"\t value of greater then 0.05, H1 Rejected & H0 Accepted")**

**print("\n\n")**

**Test Age vs Attrition:**

H0: attrited and non-attrited has no significant differences in the Age

H1: attrited and non-attrited has significant differences in the Age

Value of P is 2.9951588479067175e-30 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test BusinessTravel vs Attrition:**

H0: attrited and non-attrited has no significant differences in the BusinessTravel

H1: attrited and non-attrited has significant differences in the BusinessTravel

Value of P is 0.024442388462694316 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test Department vs Attrition:**

H0: attrited and non-attrited has no significant differences in the Department

H1: attrited and non-attrited has significant differences in the Department

Value of P is 0.0027584840572994446 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test DistanceFromHome vs Attrition:**

H0: attrited and non-attrited has no significant differences in the DistanceFromHome

H1: attrited and non-attrited has significant differences in the DistanceFromHome

Value of P is 0.4629185205822659 & decision is as following

value of greater then 0.05, H1 Rejected & H0 Accepted

**Test Education vs Attrition:**

H0: attrited and non-attrited has no significant differences in the Education

H1: attrited and non-attrited has significant differences in the Education

Value of P is 0.12035477215449608 & decision is as following

value of greater then 0.05, H1 Rejected & H0 Accepted

**Test Gender vs Attrition:**

H0: attrited and non-attrited has no significant differences in the Gender

H1: attrited and non-attrited has significant differences in the Gender

Value of P is 0.11439304195948907 & decision is as following

value of greater then 0.05, H1 Rejected & H0 Accepted

**Test JobLevel vs Attrition:**

H0: attrited and non-attrited has no significant differences in the JobLevel

H1: attrited and non-attrited has significant differences in the JobLevel

Value of P is 0.4211326530832555 & decision is as following

value of greater then 0.05, H1 Rejected & H0 Accepted

**Test JobRole vs Attrition:**

H0: attrited and non-attrited has no significant differences in the JobRole

H1: attrited and non-attrited has significant differences in the JobRole

Value of P is 0.07748676434031176 & decision is as following

value of greater then 0.05, H1 Rejected & H0 Accepted

**Test MonthlyIncome vs Attrition:**

H0: attrited and non-attrited has no significant differences in the MonthlyIncome

H1: attrited and non-attrited has significant differences in the MonthlyIncome

Value of P is 0.053577283839938566 & decision is as following

value of greater then 0.05, H1 Rejected & H0 Accepted

**Test NumCompaniesWorked vs Attrition:**

H0: attrited and non-attrited has no significant differences in the NumCompaniesWorked

H1: attrited and non-attrited has significant differences in the NumCompaniesWorked

Value of P is 0.028098156273425677 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test PercentSalaryHike vs Attrition:**

H0: attrited and non-attrited has no significant differences in the PercentSalaryHike

H1: attrited and non-attrited has significant differences in the PercentSalaryHike

Value of P is 0.018660129917539733 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test TotalWorkingYears vs Attrition:**

H0: attrited and non-attrited has no significant differences in the TotalWorkingYears

H1: attrited and non-attrited has significant differences in the TotalWorkingYears

Value of P is 1.3683394202436042e-39 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test TrainingTimesLastYear vs Attrition:**

H0: attrited and non-attrited has no significant differences in the TrainingTimesLastYear

H1: attrited and non-attrited has significant differences in the TrainingTimesLastYear

Value of P is 0.005167954938699059 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test YearsAtCompany vs Attrition:**

H0: attrited and non-attrited has no significant differences in the YearsAtCompany

H1: attrited and non-attrited has significant differences in the YearsAtCompany

Value of P is 6.047598261692858e-37 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test YearsSinceLastPromotion vs Attrition:**

H0: attrited and non-attrited has no significant differences in the YearsSinceLastPromotion

H1: attrited and non-attrited has significant differences in the YearsSinceLastPromotion

Value of P is 0.0002021180346719736 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test YearsWithCurrManager vs Attrition:**

H0: attrited and non-attrited has no significant differences in the YearsWithCurrManager

H1: attrited and non-attrited has significant differences in the YearsWithCurrManager

Value of P is 1.2365483142169853e-31 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

1. Parametric t-Test (Separate T Test)

**from scipy.stats import ttest\_ind**

**!**

**for i in num\_cols:**

**print(f"Test {i} vs Attrition: \n")**

**print(f"\tH0: attrited and non-attrited has no significant differences in the {i}")**

**print(f"\tH1: attrited and non-attrited has significant differences in the {i}")**

**stats, p = ttest\_ind(dset1\_att\_y[i],dset1\_att\_n[i])**

**print("\n\t Value of P is ",p, " & decision is as following\n")**

**if p < 0.05:**

**print(f"\t value of less then 0.05, H0 Rejected & H1 Accepted")**

**else:**

**print(f"\t value of greater then 0.05, H1 Rejected & H0 Accepted")**

**print("\n\n")**

**Test Age vs Attrition:**

H0: attrited and non-attrited has no significant differences in the Age

H1: attrited and non-attrited has significant differences in the Age

Value of P is 1.9968016158894556e-26 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test BusinessTravel vs Attrition:**

H0: attrited and non-attrited has no significant differences in the BusinessTravel

H1: attrited and non-attrited has significant differences in the BusinessTravel

Value of P is 0.9960919945450981 & decision is as following

value of greater then 0.05, H1 Rejected & H0 Accepted

**Test Department vs Attrition:**

H0: attrited and non-attrited has no significant differences in the Department

H1: attrited and non-attrited has significant differences in the Department

Value of P is 0.0013638319632118807 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test DistanceFromHome vs Attrition:**

H0: attrited and non-attrited has no significant differences in the DistanceFromHome

H1: attrited and non-attrited has significant differences in the DistanceFromHome

Value of P is 0.518286042805572 & decision is as following

value of greater then 0.05, H1 Rejected & H0 Accepted

**Test Education vs Attrition:**

H0: attrited and non-attrited has no significant differences in the Education

H1: attrited and non-attrited has significant differences in the Education

Value of P is 0.3157293177122392 & decision is as following

value of greater then 0.05, H1 Rejected & H0 Accepted

**Test Gender vs Attrition:**

H0: attrited and non-attrited has no significant differences in the Gender

H1: attrited and non-attrited has significant differences in the Gender

Value of P is 0.22881970951816322 & decision is as following

value of greater then 0.05, H1 Rejected & H0 Accepted

**Test JobLevel vs Attrition:**

H0: attrited and non-attrited has no significant differences in the JobLevel

H1: attrited and non-attrited has significant differences in the JobLevel

Value of P is 0.4945171727187496 & decision is as following

value of greater then 0.05, H1 Rejected & H0 Accepted

**Test JobRole vs Attrition:**

H0: attrited and non-attrited has no significant differences in the JobRole

H1: attrited and non-attrited has significant differences in the JobRole

Value of P is 0.08658208267572462 & decision is as following

value of greater then 0.05, H1 Rejected & H0 Accepted

**Test MonthlyIncome vs Attrition:**

H0: attrited and non-attrited has no significant differences in the MonthlyIncome

H1: attrited and non-attrited has significant differences in the MonthlyIncome

Value of P is 0.03842748490605113 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test NumCompaniesWorked vs Attrition:**

H0: attrited and non-attrited has no significant differences in the NumCompaniesWorked

H1: attrited and non-attrited has significant differences in the NumCompaniesWorked

Value of P is 0.005033438384344648 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test PercentSalaryHike vs Attrition:**

H0: attrited and non-attrited has no significant differences in the PercentSalaryHike

H1: attrited and non-attrited has significant differences in the PercentSalaryHike

Value of P is 0.03074338643339195 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test TotalWorkingYears vs Attrition:**

H0: attrited and non-attrited has no significant differences in the TotalWorkingYears

H1: attrited and non-attrited has significant differences in the TotalWorkingYears

Value of P is 5.473159751797357e-30 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test TrainingTimesLastYear vs Attrition:**

H0: attrited and non-attrited has no significant differences in the TrainingTimesLastYear

H1: attrited and non-attrited has significant differences in the TrainingTimesLastYear

Value of P is 0.0010247061915374478 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test YearsAtCompany vs Attrition:**

H0: attrited and non-attrited has no significant differences in the YearsAtCompany

H1: attrited and non-attrited has significant differences in the YearsAtCompany

Value of P is 3.163883122491456e-19 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test YearsSinceLastPromotion vs Attrition:**

H0: attrited and non-attrited has no significant differences in the YearsSinceLastPromotion

H1: attrited and non-attrited has significant differences in the YearsSinceLastPromotion

Value of P is 0.028330336189428353 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

**Test YearsWithCurrManager vs Attrition**:

H0: attrited and non-attrited has no significant differences in the YearsWithCurrManager

H1: attrited and non-attrited has significant differences in the YearsWithCurrManager

Value of P is 1.7339322652918153e-25 & decision is as following

value of less then 0.05, H0 Rejected & H1 Accepted

1. Correlation Analysis

**import pandas as pd**

**import matplotlib.pyplot as plt**

**dataset=pd.read\_csv("general\_data.csv")**

**from scipy.stats import pearsonr**

**dataset.drop\_duplicates()**

**dataset.dropna()**

**!**

**dataset.describe()**

Out[8]:

Age ... YearsWithCurrManager

count 4410.000000 ... 4410.000000

mean 36.923810 ... 4.123129

std 9.133301 ... 3.567327

min 18.000000 ... 0.000000

25% 30.000000 ... 2.000000

50% 36.000000 ... 3.000000

75% 43.000000 ... 7.000000

max 60.000000 ... 17.000000

[8 rows x 16 columns]

**dataset["Attrition"].replace(to\_replace=("No","Yes"),value=(0,1),inplace=True)**

**dataset["Gender"].replace(to\_replace=("Male","Female"),value=(0,1),inplace=True)**

* 1. Attrition and Age

**stats, p=pearsonr(dataset.Attrition, dataset.Age)**

**print (stats, p)**

**-0.15920500686577965 1.996801615886744e-26**

Null : There is no relation between employee age and attrition

Alternate: There is relation between employee age and attrition

P Less than 0.05 Null Hypothesis rejected

* 1. Attrition and Gender

**stats, p=pearsonr(dataset.Attrition, dataset.Gender)**

**print (stats, p)**

**-0.01812507887701024 0.22881970951795952**

Null : There is no relation between employee Gender and attrition

Alternate: There is relation between employee ag Gender and attrition

P greater than 0.05 Null Hypothesis accepted

* 1. Attrition and Job Level

**stats, p=pearsonr(dataset.Attrition, dataset.JobLevel)**

**print (stats, p)**

**-0.010289713287495035 0.49451717271828405**

Null : There is no relation between employee JobLevel and attrition

Alternate: There is relation between employee Job Level and attrition

P greater than 0.05 Null Hypothesis accepted

* 1. Attrition and YearsWithCurrManager

**stats, p=pearsonr(dataset.Attrition, dataset.YearsWithCurrManager)**

**print (stats, p)**

**-0.15619931590162842 1.7339322652900218e-25**

Null : There is no relation between employee YearsWithCurrManager and attrition

Alternate: There is relation between employee YearsWithCurrManager and attrition

P greater than 0.05 Null Hypothesis accepted

* 1. Attrition andDistanceFromHome

**stats, p=pearsonr(dataset.Attrition, dataset.DistanceFromHome)**

**print (stats, p)**

**-0.00973014101017966 0.5182860428050771**

Null : There is no relation between employee DistanceFromHome and attrition

Alternate: There is relation between employee DistanceFromHome and attrition

P greater than 0.05 Null Hypothesis accepted

* 1. Attrition and YearsSinceLastPromotion

**stats, p=pearsonr(dataset.Attrition, dataset.YearsSinceLastPromotion)**

**print (stats, p)**

**-0.03301877514258434 0.028330336189396753**

Null : There is no relation between employee YearsSinceLastPromotion and attrition

Alternate: There is relation between employee YearsSinceLastPromotion and attrition

P Less than 0.05 Null Hypothesis rejected

* 1. Attrition and MonthlyIncome

**stats, p=pearsonr(dataset.Attrition, dataset.MonthlyIncome)**

**print (stats, p)**

**-0.031176281698115007 0.03842748490600132**

Null : There is no relation between employee MonthlyIncome and attrition

Alternate: There is relation between employee MonthlyIncome and attrition

P Less than 0.05 Null Hypothesis rejected

* 1. Attrition and PercentSalaryHike

**stats, p=pearsonr(dataset.Attrition, dataset.PercentSalaryHike)**

**print (stats, p)**

**0.03253259489105351 0.030743386433355353**

Null : There is no relation between employee PercentSalaryHike and attrition

Alternate: There is relation between employee PercentSalaryHike and attrition

P Less than 0.05 Null Hypothesis rejected

* 1. Attrition and YearsAtCompany

**stats, p=pearsonr(dataset.Attrition, dataset.YearsAtCompany)**

**print (stats, p)**

**-0.1343922139899772 3.1638831224877484e-19**

Null : There is no relation between employee YearsAtCompany and attrition

Alternate: There is relation between employee YearsAtCompany and attrition

P greater than 0.05 Null Hypothesis accepted

* 1. Attrition and TrainingTimesLastYear

**stats, p=pearsonr(dataset.Attrition, dataset.TrainingTimesLastYear)**

**print (stats, p)**

**-0.04943057624425501 0.0010247061915362814**

Null : There is no relation between employee TotalWorkingYears and attrition

Alternate: There is relation between employee TotalWorkingYears and attrition

P Less than 0.05 Null Hypothesis rejected